Análise evolutiva de alguns indicadores de desenvolvimento em países europeus: a metodologia STATIS

Patrícia Carvalhido
Faculdade de Engenharia da Universidade do Porto, Departamento de Engenharia Civil
Adelaide Figueiredo
Faculdade de Economia da Universidade do Porto

Resumo: No presente estudo aplica-se a metodologia STATIS (*Structuration des Tableaux à Trois Indices de la Statistique*) na análise da evolução de alguns indicadores de desenvolvimento em países europeus entre 1996 e 1999. Esta metodologia permite comparar globalmente os vários quadros de dados e descrever a evolução de cada país ao longo do período de tempo considerado (através da sua trajectória em torno do "país médio") ou a evolução de cada indicador (através da sua trajectória em torno do "indicador médio"). Na análise estatística foram utilizados os programas Matlab 6.5 e Spad 5.5.

Palavras–chave: Análise em Componentes Principais, coeficiente RV, metodologia STATIS

Abstract: In the present study we apply the STATIS methodology (*Structuration des Tableaux à Trois Indices de la Statistique*) in the analysis of the evolution of some development indicators in European countries between 1996 and 1999. This methodology compares several data tables, describing the evolution of each country in that period (through its trajectory around the "average country") and the evolution of each indicator (through its trajectory around the "average indicator"). The statistical analysis was performed using Matlab 6.5 and Spad 5.5.

Keywords: Principal Component Analysis, RV-coefficient, STATIS methodology

Classificação MSC2000: 62-07; 62H25.

1 Introdução

Os dados analisados[1] dizem respeito a alguns indicadores de desenvolvimento nos anos de 1996 a 1999 em dezasseis países europeus. Nesta análise há dois aspectos a ter em conta: a evolução dos diversos países e a evolução dos indicadores no período referido. Para tal, aplicou-se a metodologia STATIS ("STATIS" e "STATIS dual") com o auxílio dos programas Matlab 6.5 e Spad 5.5., precedida de uma análise descritiva dos dados.

[1] Fonte: World Development Indicators 2004 (The World Bank Group)

2 Metodologia

A metodologia STATIS (*Structuration des Tableaux à Trois Indices de la Statistique*) foi inicialmente abordado por Escoufier (1973) e L'Hermier des Plantes (1976), e posteriormente desenvolvido por Lavit (1988, 1994). Esta metodologia não se restringe apenas à análise de um quadro de dados como é o caso da Análise em Componentes Principais (ACP), mas permite o estudo simultâneo de vários quadros de dados (quantitativos) recolhidos de uma das seguintes formas:

- T quadros de dados $(X_t)_{n\times p_t}$ recolhidos em diferentes "ocasiões" sobre n indivíduos; as variáveis podem diferir ao longo dos quadros; neste caso a metodologia é caracterizado por T estudos (X_t, Q_t, D), $t = 1, \ldots, T$, em que Q_t é a métrica no espaço dos indivíduos e D é a métrica no espaço das variáveis, i.e., a matriz diagonal dos pesos dos indivíduos;

- as mesmas variáveis são descritas ao longo de T quadros de dados $(X_t)_{n_t\times p}$ recolhidos em diferentes "ocasiões" sobre indivíduos que podem diferir ao longo dos quadros; neste caso a metodologia é caracterizado por T estudos (X_t, Q, D_t), $t = 1, \ldots, T$;

A cada uma destas situações corresponde uma estratégia diferente: a primeira evidencia as relações entre os indivíduos ("STATIS") e a segunda privilegia as relações entre as variáveis ("STATIS dual"). Para estudar quadros que "cruzem" os mesmos indivíduos com as mesmas variáveis é conveniente aplicar as duas metodologias. Estas compreendem as seguintes etapas fundamentais:

I) interestrutura: comparação global dos quadros de dados;

II) compromisso: "resumo" da informação contida nos T quadros numa matriz média;

III) intraestrutura: permite destacar os indivíduos ("STATIS") ou as variáveis ("STATIS dual") responsáveis pelas semelhanças (ou diferenças) entre os quadros; a partir da imagem compromisso traçam-se as trajectórias que descrevem o comportamento evolutivo de cada indivíduo (ou variável).

2.1 Metodologia "STATIS"

Na primeira etapa (**interestrutura**) é necessário definir um objecto representativo de cada quadro, que não é mais do que a matriz dos produtos escalares entre os indivíduos: $W_t = X_t Q_t X_t'$ $(t = 1, \ldots, T)$. De seguida, interessa estabelecer uma distância entre os objectos W_t; para tal é necessário definir um produto escalar entre os objectos:

$$\langle W_t | W_{t'} \rangle_{HS} = tr\left(DW_t DW_{t'}\right), \tag{1}$$

denominado produto escalar de Hilbert-Schmidt. No caso dos objectos normados o produto escalar entre estes corresponde ao coeficiente de correlação vectorial entre objectos, ou coeficiente RV, proposto por Robert e Escoufier (1976):

$$RV\left(t,t'\right)=\left\langle \frac{W_t}{\|W_t\|_{HS}}\,\middle|\,\frac{W_{t'}}{\|W_{t'}\|_{HS}}\right\rangle_{HS}=\frac{tr\left(DW_tDW_{t'}\right)}{\sqrt{tr\left(DW_t\right)^2}\sqrt{tr\left(DW_{t'}\right)^2}}. \tag{2}$$

O coeficiente RV permite obter a distância entre dois objectos normados uma vez que

$$d_{HS}\left(\frac{W_t}{\|W_t\|_{HS}},\frac{W_{t'}}{\|W_{t'}\|_{HS}}\right)=\left\|\frac{W_t}{\|W_t\|_{HS}}-\frac{W_{t'}}{\|W_{t'}\|_{HS}}\right\|_{HS}=\sqrt{2-2RV\left(t,t'\right)}. \tag{3}$$

Seja S a matriz dos coeficientes RV e Δ a matriz diagonal dos pesos π_t atribuídos aos quadros X_t $(t=1,\ldots,T)$. A imagem euclideana dos objectos é obtida efectuando uma ACP à matriz S. Sendo τ_t o t-ésimo valor próprio de $S\Delta$ associado ao vector próprio γ_t $(t=1,\ldots,T)$, as coordenadas dos pontos A_t (associados aos objectos W_t) no i-ésimo eixo são as componentes do vector $\sqrt{\tau_i}\gamma_i$. Se os pesos π_t forem iguais basta diagonalizar a matriz S. Esta representação é idêntica à representação das variáveis na ACP.

Por outro lado, é possível efectuar a imagem euclideana centrada dos objectos (normados). Esta permite a visualização das oposições e proximidades entre estes. Neste caso a matriz em questão, centrada pelas linhas e colunas, é

$$\widetilde{S}=\left(I_T-\mathbf{1}\mathbf{1}'\Delta\right)S\left(I_T-\Delta\mathbf{1}\mathbf{1}'\right), \tag{4}$$

em que I_T é a matriz identidade de ordem T e $\mathbf{1}$ é o vector de dimensão T com todas as componentes iguais a 1. A diagonalização da matriz $\widetilde{S}\Delta$ ou $\widetilde{S}$ (no caso de os pesos atribuídos aos objectos serem todos iguais) permite a representação dos objectos, análoga à representação dos indivíduos na ACP.

Se, de facto, houver semelhança entre os objectos, é conveniente determinar o **compromisso** W, i.e., o objecto "resumo" de todos os outros objectos e que corresponde simplesmente a uma média ponderada destes (normados):

$$W=\sum_{t=1}^{T}\alpha_t\frac{W_t}{\|W_t\|_{HS}}. \tag{5}$$

Os coeficientes α_t em (5) são tais que $\|W\|_{HS}=\sum_{t=1}^{T}\pi_t\|W_t\|_{HS}$; no caso dos objectos normados tem-se $\|W\|_{HS}=1$ e

$$\alpha_t=\frac{1}{\sqrt{\tau_1}}\pi_t\gamma_1^{(t)} \tag{6}$$

em que $\gamma_1^{(t)}$ é a t-ésima coordenada de γ_1 (vector próprio de $S\Delta$ associado ao maior valor próprio τ_1).

Aplicando uma ACP à matriz W, obtém-se a imagem euclideana do compromisso. Se μ_i é o valor próprio da matriz WD associado ao vector próprio ε_i D-ortonormado ($i = 1, \ldots, n$) e $B_1, B_2, \ldots, B_n$ os pontos associados aos indivíduos na imagem euclideana do compromisso, as coordenadas destes sobre o eixo k são as componentes do vector $\sqrt{\mu_k}\varepsilon_k$. Tal como na ACP, só depois de analisar as correlações das variáveis com os eixos do compromisso, é que se interpretam as posições compromisso.

Uma vez representado o compromisso, é necessário explicar as oposições entre objectos ao nível individual (**intraestrutura**). Para tal considera-se a decomposição do quadrado da distância entre objectos em contribuições de indivíduos:

$$\frac{D(i,i)\sum_{j=1}^{n} D(j,j)\left[W_t(i,j) - W_{t'}(i,j)\right]^2}{\left[d_{HS}(W_t, W_{t'})\right]^2}, \tag{7}$$

ou então a decomposição da soma dos quadrados das distâncias entre pares de objectos em contribuições de indivíduos.

De forma a visualizar graficamente estes desvios individuais traçam-se as diferentes posições de cada indivíduo para cada objecto na imagem euclideana do compromisso, designadas por "trajectórias". As coordenadas dos pontos $B_1^t, \ldots, B_n^t$ $(t = 1, \ldots, T)$ no eixo k são $\frac{1}{\sqrt{\mu_k}} W_t D \varepsilon_k$. O ponto compromisso B_i é uma média dos pontos $B_i^1, \ldots, B_i^T$ ponderada pelos coeficientes $\alpha_1, \ldots, \alpha_T$ definidos em (6).

2.2 Métodologia "STATIS dual"

Neste caso interessa saber quais as variáveis responsáveis pelas oposições e semelhanças entre os vários quadros. A matriz dos produtos escalares entre as variáveis, $V_t = X_t' D_t X_t$ $(t = 1, \ldots, T)$, constitui o objecto representativo de cada quadro e é designada por matriz de variâncias e covariâncias (ou matriz das correlações se as variáveis estiverem centradas e reduzidas). O produto escalar utilizado para induzir uma distância entre os objectos V_t e $V_{t'}$ é análogo a (1):

$$\langle V_t | V_{t'} \rangle_{HS} = tr\left(Q V_t Q V_{t'}\right). \tag{8}$$

Seja Z a matriz dos produtos escalares entre os objectos. A diagonalização de $Z\Delta$ permite a obtenção da imagem euclideana. Havendo semelhança entre os objectos, é conveniente estabelecer o compromisso:

$$V = \frac{\sum_{t=1}^{T} \beta_t V_t}{\sum_{t=1}^{T} \beta_t}, \tag{9}$$

cuja representação gráfica é obtida à custa da diagonalização de VQ. Sobre esta representação "constroem-se" as trajectórias de cada variável.

3 Aplicação

Os dados em questão[1] referem-se ao período compreendido entre 1996 e 1999. Dispõe-se então de quatro quadros de dados X_{96}, X_{97}, X_{98} e X_{99}, cada um com dezasseis linhas e sete colunas, em que cada linha corresponde a um indivíduo e cada coluna a uma variável. Não há dados omissos. Os indivíduos estudados são a Alemanha (ALE), Bélgica (BEL), Bulgária (BUL), Dinamarca (DIN), Eslovénia (ESL), Espanha (ESP), Finlândia (FIN), França (FRA), Hungria (HUN), Irlanda (IRL), Polónia (POL), Portugal (POR), Roménia (ROM), Reino Unido (RU), Suécia (SUE) e Turquia (TUR). As variáveis em causa são assinantes de TV por cabo[2] (TVC), utilizadores de telemóveis[3] (TM), computadores pessoais[3] (CP), linhas telefónicas[3] (LIT), aparelhos de TV[3] (ATV), consumo de energia eléctrica[3] (CEE) e artigos científicos e técnicos publicados (ART). Os países são afectados pelos mesmos pesos e dada a diferença de dispersões entre variáveis estas foram centradas e reduzidas em cada quadro. Os estudos em questão são $\left(X_t, Q_t, \frac{1}{16} \cdot I_{16}\right)$ em que Q_t é a matriz diagonal dos inversos das variâncias das variáveis de X_t, $t = 96, 97, 98, 99$. Os estudos têm todos o mesmo peso logo $\Delta = \frac{1}{4} \cdot I_4$.

Pode-se então analisar os objectos W_t, estudando a evolução das distâncias entre países, ou os objectos V_t, caracterizados pelas relações existentes entre as variáveis.

3.1 Estudo das distâncias individuais: metodologia "STATIS"

A matriz dos coeficientes RV bem como as figuras 1 e 2 evidenciam a forte semelhança existente entre os vários anos, que se acentua entre anos consecutivos. Destes, aqueles que estão menos "próximos" são os anos de 1998 e 1999. Sendo os coeficientes RV elevados, sem excepção, significa que as distâncias entre objectos normados são pequenas e que a evolução dos indivíduos é estável de um ano para o outro.

Tabela 1: Matriz dos coeficientes RV.

	1996	1997	1998	1999
1996	1			
1997	0.995	1		
1998	0.991	0.996	1	
1999	0.980	0.984	0.990	1

Também é de salientar a evolução cronológica regular fundamentalmente no período de 1996 a 1998. Os coeficientes α_t do objecto compromisso[4] são

[2] em milhares
[3] em kwh per capita
[4] Os objectos W_t foram normados.

quase iguais (≈ 0.251), como seria de esperar pois traduzem os coeficientes RV bastante elevados.

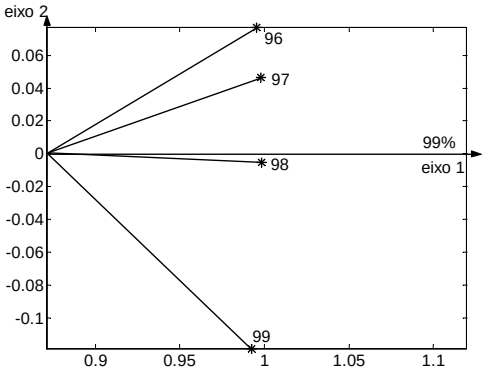

Figura 1: Imagem euclideana.

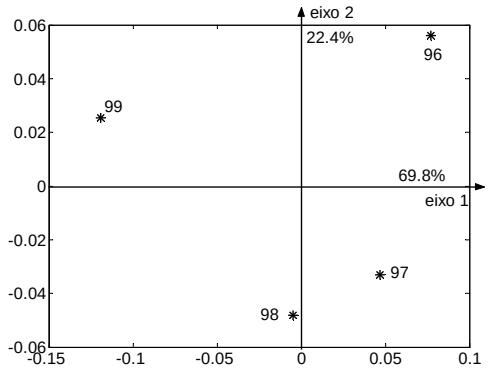

Figura 2: Imagem euclideana centrada.

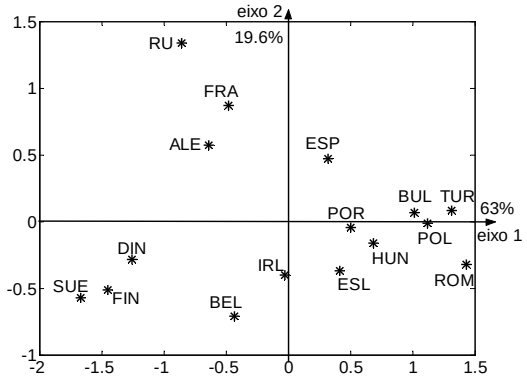

Figura 3: Posições compromisso.

Para interpretar as posições compromisso dos países (figura 3) determinam-se as correlações das variáveis com os dois primeiros eixos do compromisso. Saliente-se que, para cada variável estas correlações variam muito pouco ao longo do período considerado (tabela 2), o que significa que há pouca evolução por parte destas. As variáveis mais correlacionadas com o 1º eixo (63% da inércia total) são CEE, TM, CP, LIT e ATV. A valores mais elevados destas variáveis na Dinamarca, Finlândia e Suécia opõem-se valores mais baixos na

Bulgária, Hungria, Polónia, Roménia e Turquia. A variável mais correlacionada com o 2º eixo (19.6% da inércia total) é ART e a valores mais elevados desta por parte do Reino Unido e da França opõe-se valores mais baixos na Irlanda. Note-se que os países referidos têm elevadas contribuições relativas logo a sua qualidade de representação na imagem euclideana do compromisso é razoável.

Tabela 2: Correlações das variáveis com os eixos do compromisso.

	TVC		CEE		TM		CP	
	1º eixo	*2º eixo*	*1º eixo*	*2º eixo*	*1º eixo*	*2º eixo*	*1º eixo*	*2º eixo*
1996	-0.523	-0.554	-0.837	-0.265	-0.833	-0.221	-0.935	-0.098
1997	-0.517	-0.568	-0.836	-0.274	-0.832	-0.250	-0.924	-0.111
1998	-0.512	-0.580	-0.835	-0.274	-0.831	-0.195	-0.927	-0.086
1999	-0.472	-0.584	-0.838	-0.269	-0.834	-0.066	-0.910	-0.081

	ART		LIT		ATV	
	1º eixo	*2º eixo*	*1º eixo*	*2º eixo*	*1º eixo*	*2º eixo*
1996	-0.407	0.801	-0.927	0.106	-0.770	0.494
1997	-0.405	0.797	-0.929	0.097	-0.782	0.469
1998	-0.400	0.796	-0.927	0.100	-0.792	0.400
1999	-0.398	0.803	-0.926	0.086	-0.781	0.391

Interessa agora destacar os países responsáveis pelas diferenças entre os vários anos. A decomposição da soma dos quadrados das distâncias entre pares de objectos normados permite distinguir os países que mais contribuem para a diferença de estrutura no período considerado: Finlândia (17%), Suécia (11.1%), Dinamarca (9.3%), Bulgária (8.4%), Turquia (8.4%), Eslovénia (7.6%), Portugal (7.2%) e Roménia (6.5%). Os países que têm contribuições mais reduzidas são a Alemanha (0.9%), Hungria (2.1%) e a Bélgica (2.2%).

Considerando os anos de 1996 e 1999, os países que mais contribuem para a diferença entre estes anos (através da decomposição do quadrado da distância entre objectos normados) são a Suécia (13.8%), a Finlândia (11.1%), a Bulgária (10%), a Eslovénia (9.6%) e Portugal (8.3%). No período 1996/1997 as contribuições mais elevadas são da Finlândia (26.3%), Dinamarca (14.8%) e Suécia (9.5%), no período 1996/1998 a Finlândia (14.4%), Portugal (13.9%), Suécia (12.6%) e Bulgária (8.9%), entre 1997 e 1998 a Dinamarca (22.5%), Portugal (12.5%), a Finlândia (10.7%) e o Reino Unido (9.4%), entre 1997 e 1999, a Finlândia (21.8%), a Suécia (10.1%) e a Dinamarca (10.5%) e no período de 1998 a 1999 os países que mais contribuem para as diferenças consideradas são a Finlândia (20.8%), a Turquia (12.4%) e a Eslovénia (10.9%).

A representação das trajectórias deve agora ser feita, mas apenas naqueles indivíduos que estão bem representados no compromisso, i.e., aqueles que têm elevadas contribuições relativas (superiores a 50%). Como as contribuições

relativas ou são elevadas para o primeiro eixo ou para o segundo optou-se pela representação gráfica separada.

As figuras 4 e 5 mostram que todos os países têm trajectórias bastante homogéneas ao longo do tempo, sendo de destacar a Polónia e a Hungria, cuja evolução segue a evolução média, ou seja, para cada variável (correlacionada com o 1º eixo) o desvio entre o valor desta para estes países e a média é regular de um ano para o outro.

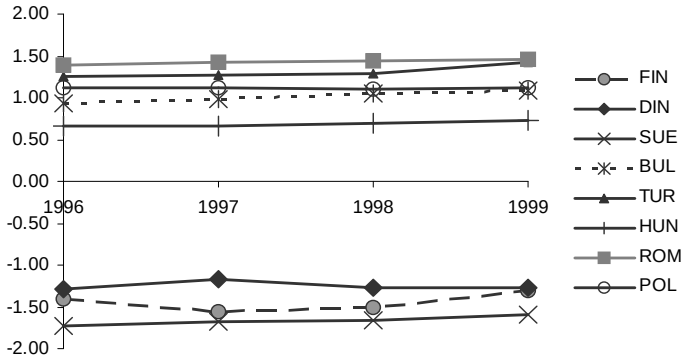

Figura 4: Trajectórias individuais em relação ao 1º eixo do compromisso.

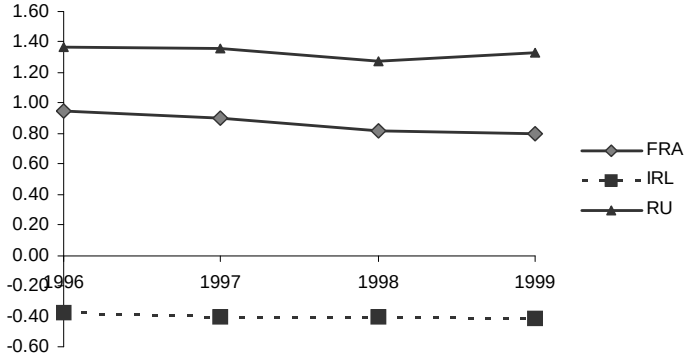

Figura 5: Trajectórias individuais em relação ao 2º eixo do compromisso.

Um aumento dos valores na figura 4 evidencia uma diminuição dos valores

das variáveis correlacionadas com este eixo em relação à evolução média, por outro lado, uma diminuição dos valores na figura 5 significa que há uma diminuição dos valores das variáveis correlacionadas com o 2º eixo.

Tal como já se havia constatado na imagem euclideana do compromisso, há uma oposição nítida entre a Dinamarca, a Finlândia e a Suécia por um lado e a Bulgária, Hungria, Polónia, Roménia e Turquia por outro (figura 4). Também é nítida a oposição existente entre a França e o Reino Unido por um lado e a Irlanda por outro (figura 5).

3.2 Estudo das correlações entre variáveis: metodologia "STATIS dual"

A interestrutura realizada sobre os objectos V_t fornece os produtos escalares (e, consequentemente, as distâncias) entre matrizes de correlações. A matriz dos coeficientes RV tem todos os seus elementos entre 0.9964 ($\approx RV(97,99)$) e 0.9995 ($\approx RV(96,98)$), o que significa que há estabilidade nas variáveis durante o período estudado. Essa estabilidade é reforçada pela determinação dos coeficientes β_t do objecto compromisso, que são extremamente semelhantes ($\beta_{96} = 0.252, \beta_{97} = 0.251, \beta_{98} = 0.250, \beta_{99} = 0.246$), fazendo com que o compromisso seja quase uma média aritmética dos objectos, ou seja, este é um "reflexo" bastante razoável das várias matrizes de correlações. O objecto mais "próximo" do objecto compromisso é V_{98} ($d_{HS}(V, V_{98}) = 0.1$) seguido de V_{97} (0.155), V_{96} (0.160) e por último V_{99} (0.267) como se pode confrontar com a figura 6. [5]

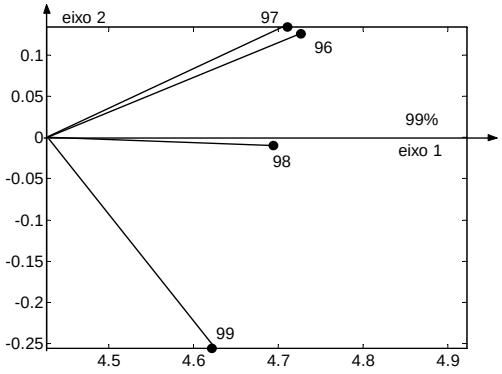

Figura 6: Imagem euclideana.

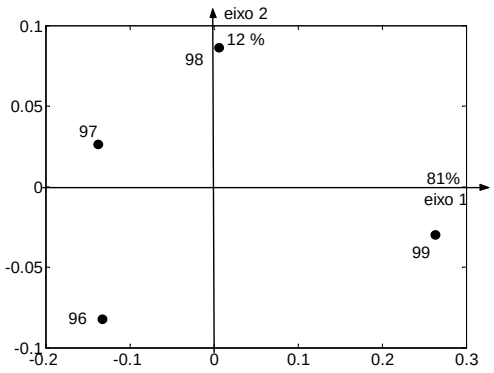

Figura 7: Imagem euclideana centrada.

Observando a figura 8, verifica-se que as variáveis situam-se do lado positivo do 1º eixo, com a excepção da variável ART e da variável TVC, cuja qualidade de

[5] Note-se que V situa-se no 1º eixo da imagem euclideana.

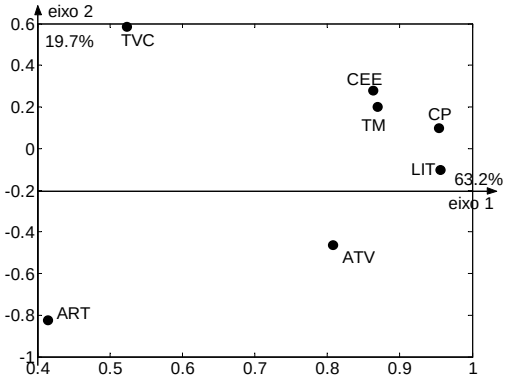

Figura 8: Posições compromisso.

representação é pouco satisfatória (cfr. tabela 2). Há uma grande proximidade entre as variáveis CEE e TM, e entre CP e LIT.

As trajectórias de cada uma das variáveis encontram-se reunidas praticamente em cima das respectivas posições compromisso, o que leva a crer que há uma forte estabilidade por parte destas neste período. A decomposição da soma dos quadrados das distâncias entre pares de objectos evidenciou as variáveis TM (28%), TVC (23%) e ART (18.3%) como sendo aquelas que mais contribuiram para as diferenças existentes no período estudado.

4 Conclusões

De entre as duas metodologias estudadas, aquela que revelou resultados mais interessantes foi a metodologia "STATIS", uma vez que se registaram evoluções mais significativas por parte dos países que dos indicadores de desenvolvimento entre 1996 e 1999.

Há grupos de países que se distinguem, tais como o grupo da Dinamarca, Finlândia e Suécia e o grupo da Bulgária, Hungria, Polónia, Roménia e Turquia. Os primeiros apresentam valores de CEE, TM, CP, LIT e ATV mais elevados por oposição a valores mais baixos por parte destes últimos. A França e o Reino Unido apresentam valores mais elevados no que diz respeito à variável ART, por oposição à Irlanda.

A Alemanha é o país mais estável ao longo do período considerado. Este facto é comprovado pela decomposição do quadrado da distância entre objectos W_t, uma vez que é o país com contribuições mais reduzidas.

A Finlândia, a Suécia e a Dinamarca são aqueles que mais contribuiram para as diferenças de estrutura nos anos em causa.

As diferenças mais significativas entre anos consecutivos registadas em Portugal verificaram-se no período 1997/1998.

Há uma diferença nítida por parte de todas as variáveis consideradas em relação à variável ART, tal como seria de esperar uma vez que esta variável é de natureza bastante diferente das outras.

As variáveis CP e LIT estão fortemente correlacionadas, assim como CEE e TM.

A metodologia STATIS revelou-se profícua para este conjunto de dados, uma vez que analisa a evolução decorrida ao longo de um período de tempo, não só das variáveis, bem como dos indivíduos em causa. No entanto, tem algumas limitações tais como o compromisso ser influenciado pelos objectos que possuem normas mais elevadas e, no caso de se utilizarem os objectos normados, corre-se o risco de não se dar relevância às diferenças de estrutura entre quadros. Em Dazy & Barzic (1996) são apresentados outros métodos de análise evolutiva de quadros de dados.

Referências

[1] Dazy, F., Le Barzic, J. (1996). *L'Analyse des Données Évolutives, methodes et applications*. Technip.

[2] Escoufier, Y. (1973). Le traitement de variables vectorielles. *Biometrics*, Vol. 29, n^{o} 4, p. 751-760.

[3] Lavit, C. (1988). *Analyse Conjointe de Tableaux Quantitatives*. Collection Méthodes+Programmes, Masson.

[4] Lavit, C., Escoufier, Y., Sabatier, R. e Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics and Data Analysis*, Vol. 18, p. 97-119.

[5] L'Hermier des Plantes, H. (1976). *Structuration des Tableaux à Trois Indices de la Statistique*. Thèse de 3 ème cycle. Université de Montpellier II.

[6] Robert, P., Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: the RV-coefficient. *Applied Statistics*, Vol. 25, n^{o} 3, p. 257-265.